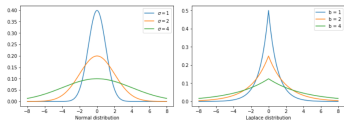


## Regularization Bayesian Priors



- RRM is same as MAP in Bayes
- Gaussian/Laplace prior corresponds to  $L_2/L_1$  penalty

- RRM is same as MAP in Bayes
- Gaussian/Laplace prior corresponds to  $L_2/L_1$  penalty

# RRM VS. BAYES

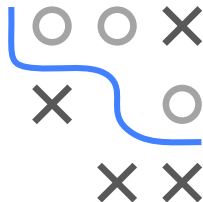
We already created a link between max. likelihood estimation and ERM.

Now we will generalize this for RRM.

Assume we have a parameterized distribution  $p(y|\theta, \mathbf{x})$  for our data and a prior  $q(\theta)$  over our param space, all in Bayesian framework.

From Bayes theorem:

$$p(\theta|\mathbf{x}, y) = \frac{p(y|\theta, \mathbf{x})q(\theta)}{p(y|\mathbf{x})} \propto p(y|\theta, \mathbf{x})q(\theta)$$



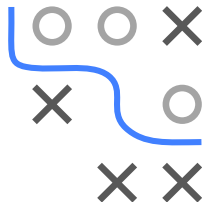
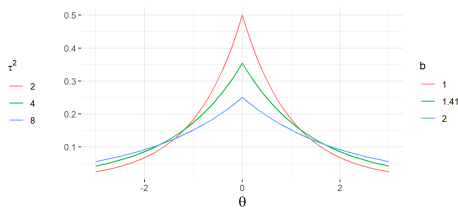
## RRM VS. BAYES / 2

The maximum a posteriori (MAP) estimator of  $\theta$  is now the minimizer of

$$-\log p(y | \theta, \mathbf{x}) - \log q(\theta).$$

- Again, we identify the loss  $L(y, f(\mathbf{x} | \theta))$  with  $-\log(p(y|\theta, \mathbf{x}))$ .
- If  $q(\theta)$  is constant (i.e., we used a uniform, non-informative prior), the second term is irrelevant and we arrive at ERM.
- If not, we can identify  $J(\theta) \propto -\log(q(\theta))$ , i.e., the log-prior corresponds to the regularizer, and the additional  $\lambda$ , which controls the strength of our penalty, usually influences the peakedness / inverse variance / strength of our prior.





- L2 regularization corresponds to a zero-mean Gaussian prior with constant variance on our parameters:  $\theta_j \sim \mathcal{N}(0, \tau^2)$
- L1 corresponds to a zero-mean Laplace prior:  $\theta_j \sim \text{Laplace}(0, b)$ .  $\text{Laplace}(\mu, b)$  has density  $\frac{1}{2b} \exp(-\frac{|\mu - x|}{b})$ , with scale parameter  $b$ , mean  $\mu$  and variance  $2b^2$ .
- In both cases, regularization strength increases as variance of prior decreases: more prior mass concentrated around 0 encourages shrinkage.
- Elastic-net regularization corresponds to a compromise between Gaussian and Laplacian priors

# EXAMPLE: BAYESIAN L2 REGULARIZATION

We can easily see the equivalence of  $L2$  regularization and a Gaussian prior:

- Gaussian prior  $\mathcal{N}_d(\mathbf{0}, \text{diag}(\tau^2))$  with uncorrelated components for  $\boldsymbol{\theta}$ :

$$q(\boldsymbol{\theta}) = \prod_{j=1}^d \phi_{0, \tau^2}(\theta_j) = (2\pi\tau^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\tau^2} \sum_{j=1}^d \theta_j^2\right)$$

- MAP:

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\text{MAP}} &= \arg \min_{\boldsymbol{\theta}} (-\log p(y | \boldsymbol{\theta}, \mathbf{x}) - \log q(\boldsymbol{\theta})) \\ &= \arg \min_{\boldsymbol{\theta}} \left( -\log p(y | \boldsymbol{\theta}, \mathbf{x}) + \frac{d}{2} \log(2\pi\tau^2) + \frac{1}{2\tau^2} \sum_{j=1}^d \theta_j^2 \right) \\ &= \arg \min_{\boldsymbol{\theta}} \left( -\log p(y | \boldsymbol{\theta}, \mathbf{x}) + \frac{1}{2\tau^2} \|\boldsymbol{\theta}\|_2^2 \right)\end{aligned}$$

- We see how the inverse variance (precision)  $1/\tau^2$  controls shrinkage



# EXAMPLE: BAYESIAN L2 REGULARIZATION / 2

- DGP  $y = \theta + \varepsilon$  where  $\varepsilon \sim \mathcal{N}(0, 1)$  and  $\theta = 1$ ;  
with Gaussian prior on  $\theta$ , so  $\mathcal{N}(0, \tau^2)$  for  $\tau \in \{0.25, 0.5, 2\}$
- For  $n = 20$ , posterior of  $\theta$  and MAP can be calculated analytically
- Plotting the  $L_2$  regularized empirical risk  $\mathcal{R}_{\text{reg}}(\theta) = \sum_{i=1}^n (y_i - \theta)^2 + \lambda \theta^2$   
with  $\lambda = 1/\tau^2$  shows that ridge solution is identical with MAP

